

DNA Sequence Evolution Simulation and Phylogeny Building with Pen and Paper

Instructor's Guide

Overview: This set of activities is designed to promote understanding of evolution of DNA sequences and phylogenetic analysis. This guide includes instructions for using the activities individually or as a linked set of exercises.

Activity 1: Interpreting Phylogenetic Trees – The instructor will introduce or review how to interpret phylogenetic trees. Students can complete an optional worksheet.

Activity 2: Simulating DNA Sequence Evolution – Groups of students will use dice (4-sided and 10-sided) to simulate accumulation of mutations in descendants of an ancestral DNA sequence as they diverge according to a given phylogenetic tree.

Activity 3: Mapping Simulated DNA Sequences to a Known Phylogeny – Groups of students will exchange the simulated DNA sequences and try to reassign them to the tips of the phylogenetic tree.

Activity 4: Building a Phylogenetic Tree from DNA Sequences – Students will use a hierarchical clustering method to reconstruct a phylogenetic tree from their simulated DNA sequences.

Length: 1-2 hour class periods, depending on homework and discussion

Intended Participants: High school and/or middle school students.

Materials: At least two groups of students are needed. Each group of students will require two polyhedral dice (one 4-sided and one 10-sided), although it is possible to share a single set of dice. Each group of students will need a pair of scissors, although it is also possible to share these. Student handouts are attached here, and an electronic version in pdf format is available by email (ntperna@wisc.edu).

Authorship and Use: The UPGMA portion was inspired by a lesson <http://csunplugged.org/wp-content/uploads/2014/12/PhylogeneticsUnplugged.pdf> developed by Tru Women in Computer Science (TWiCS) <http://twics.truman.edu> Truman State University, Kirksville, Missouri, USA (Mariya Davidkova Amy McNabb Molly Smith Julia Stefani Michelle VanKleeck Allie Wehrman and Jon Beck) and contributed to csunplugged.org by Katrin and Jim Becker, Mount Royal College; Calgary, Alberta, Canada. The dice-based simulation, instructor's guide and student handouts were created by Nicole T. Perna and Jeremy D. Glasner from the J.F. Crow Institute for the Study of Evolution at the University of Wisconsin – Madison **for non-commercial purposes only**. CSUnplugged and this exercise are distributed under a Creative Commons BY-NC-SA License, which makes it easy to copy, adapt and share <https://creativecommons.org/licenses/by-nc-sa/3.0/>.

DNA Sequence Evolution Simulation and Phylogeny Building with Pen and Paper

Instructor's Guide

Learning Objectives

- Students will be able to correctly interpret relationships depicted in a phylogenetic tree (Activity 1).
 - Students will understand how mutations accumulate in DNA sequences over evolutionary time (Activity 2).
 - Students will appreciate the role of simulation and modeling in biology and bioinformatics (Activity 2).
 - Students will recognize the relationship between expected number of differences among homologous sequences and the evolutionary path in the tree (Activity 3 and 4).
 - Students will experience one approach that scientists use to infer the evolutionary relationship between species or homologous genes (Activity 4).
-

NGSS Connections

- NGSS Core Disciplinary Ideas

 - LS-3A Inheritance of Traits
 - LS-3B Variation of Traits
 - LS-4A Common Ancestry and Diversity
 - NGSS Practices

 - Developing and Using Models
 - Analyzing and Interpreting Data
 - Using Mathematics and Computational Thinking
-

Instructor's Guide - Lesson Plan

➤ **Activity 1: Interpreting Phylogenetic Trees**

Discuss phylogenetic trees with the students. Explain that we use phylogenetic trees to depict a model of evolutionary relationships among species or genes. Emphasize the concept of common ancestry. Show an example of a tree (***Student Handout – Phylogenetic Tree***) and describe that the tips represent species or genes that are found on earth today and the internal nodes represent ancestral organisms or sequences.

Tip: Interpreting phylogenetic trees is often referred to as “Tree Thinking” and there are many resources available that you can use to supplement this part of the activity. For example, <http://www.tree-thinking.org/> provides a brief introduction to the topic, as well as more extensive material on common misconceptions, and several assessment tools you may find useful. The “Evolution 101” web site has phylogeny section that you could incorporate into your lesson http://evolution.berkeley.edu/evolibrary/article/evo_03. You might consider assigning it as homework. We estimate that it takes a half hour to work through this material.

Here, using the tree in ***Student Handout – Phylogenetic Tree*** as an example links learning to interpret trees to the portion of the exercise where students simulate DNA sequence evolution using the same tree. You may want to note that although we have added a timescale to this tree, time is not an inherent aspect of phylogenetic trees. Rather, timescales must be added based on additional knowledge, such as fossil evidence for age of an ancestral type. Given the timescale provided here, students can be encouraged to interpret the relationship between branch length and time since the species/genes at the tips last shared a common ancestor.

An optional worksheet is provided as a short assessment tool (***Tree Thinking Worksheet***). The correct answers are 1. a, 2. b, 3. a, 4. c, 5. a.

Instructor's Guide - Lesson Plan

➤ **Activity 2: Simulating DNA Sequence Evolution**

In this part of the exercise, groups of students will simulate evolution of DNA sequences. A basic understanding of the structure of DNA is useful. The brief description in ***Student Handout – DNA Sequence Evolution Simulation*** is adequate to complete this activity or will serve as a reminder if you have already covered the structure and replication of DNA. The handout introduction also includes an overview of the role of simulation in biological sciences.

Here is how this simulation works: Students will start with an ancestral DNA sequence 10 bases long and a phylogenetic tree that represents the evolutionary history of the species carrying that DNA sequence. Our simulation model assumes that a 10 base DNA sequence will experience one mutation every million years, each base in the sequence is equally likely to change, and each type of mutation is equally likely. Along each branch of the tree in ***DNA Sequence Evolution Worksheet 1***, the students will introduce mutations into the DNA sequence according to this model. To decide where in the DNA sequence, and what kind of mutation, they will roll two dice. One of the dice has 10 sides, and the number they roll will determine which of the bases (1 through 10) in the DNA sequence will be mutated. The other die has 4 sides, and the number they roll will determine the type of mutation: If they roll a 1, the base changes to A. Roll a 2, the base changes to C. Roll a 3, the base changes to G. Roll a 4, the base changes to T. Sometimes the “mutation” will result in no change at all. For example, an A will “change” to an A.

Tips: Groups of 3 students works well. One student can roll the 10-sided die, one can roll the 4-sided die, and one can record the mutated DNA sequences. Be sure the students know how to read the dice. Rolling a “0” on the 10-sided die is equivalent to rolling a “10”. Read the number at the top of the 4-sided die (it will be the same for all three visible sides). You may want to remind the students to pay attention to the branch lengths and time on the tree. A DNA sequence evolving along a branch that spans 2 million years of time should experience 2 mutations instead of 1.

Optional discussion: The assumptions of the simulation model are not explicit in the student handouts and worksheet. After they have completed the simulation, you might consider challenging them to think about the assumptions and the extent to which this model accurately represents the true process of DNA sequence evolution. Are all bases really equally likely to experience a mutation? Are all types of mutation equally common? Emphasize that all models and simulations make assumptions about the underlying process and although they may not be totally realistic, models can still be useful because they make predictions that can be tested with real data.

Instructor's Guide - Lesson Plan

➤ **Activity 3: Mapping Simulated DNA Sequence to a Known Phylogeny**

In this portion of the exercise, groups of students will exchange DNA sequences evolved in their simulation (**Activity 2**) and try to assign them back to the correct place on the tips of the trees. This is an active learning exercise intended to develop a better sense of how differences in DNA sequences found in organisms on earth today relates to the topology (or structure) of the phylogeny that reflects their evolutionary history.

Procedure: Have each group copy the DNA sequences from the tips of the tree on their **DNA Sequence Evolution Worksheet 1** to the corresponding spots on **DNA Sequence Worksheet 2**. They will cut these into strips to exchange with another group. BEFORE they begin cutting, they need to code the sequence names so the other group won't know the correct order! To do this, they should re-label the sequences currently called (A, B, C, D and E) as (F, G, H, I, and J) in a different order. Make sure they copy the coded labels into both blanks. One set stays with the original tree to keep a record of the correct answer. The other set of labels goes with the sequences.

Next, they should cut along the dotted lines and exchange sequences between groups. Then, have them try to map the sequences back onto the starting tree. After 5 or 10 minutes, ask them to make a final selection. Ask them about the process they used to make their decision. They can compare their own results to the correct answer from the other group. If they got the wrong answer, ask them to think of possible explanations.

Tips: It can be quite hard to get the correct answer, especially if some of the simulated mutations occurred at the same position more than once within a lineage, or at the same position in different lineages.

Note: The reason they need to make two cuts between each sequence is that we discovered that students are otherwise able to align the sequences based on the cut patterns alone.

DNA Sequence Evolution Simulation and Phylogeny Building with Pen and Paper

Instructor's Guide - Lesson Plan

➤ Activity 4: Building a Phylogenetic Tree from DNA Sequences

The final portion of this exercise involves building a tree from the sequences evolved in Activity 2 using a pen, paper and some straightforward arithmetic. You may choose to have the students do it independently after you illustrate the approach, or you can walk them through the exercise. You will use a method called **UPGMA**, which stands for **Unweighted Pair Group Method with Arithmetic Mean**.

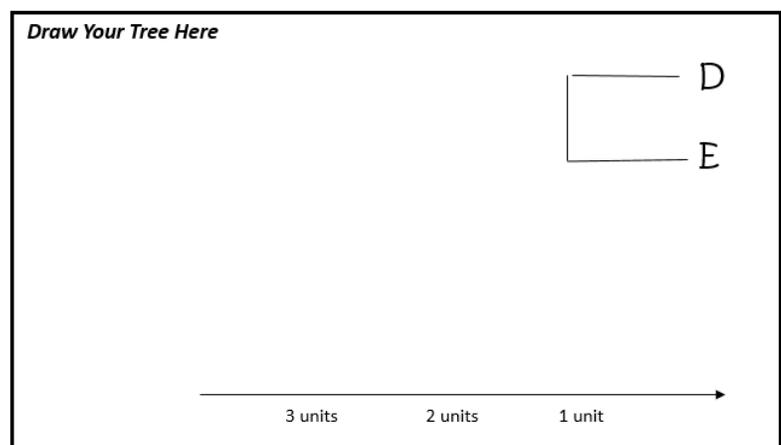
Procedure:

Round 1: The first step is to count the number of differences between each pair of evolved sequences, and enter them into the first "Distance Matrix" on the **UPGMA Worksheet**. To make it easier, line up pairs of the strips of paper containing the DNA sequences. First, compare A to B. Then, A to C, A to D, A to E, B to C, B to D, B to E, C to D, C to E, and finally D to E. Notice that comparing A to B is the same as comparing B to A, so we only need to use the lower triangle of the matrix. It should look something like the one on the right.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | - | | | | |
| B | 4 | - | | | |
| C | 6 | 6 | - | | |
| D | 6 | 6 | 4 | - | |
| E | 6 | 6 | 4 | 2 | - |

Find the smallest value in the matrix. That is the distance between the two most closely related sequences, which corresponds to the total branch length separating the two sequences in the tree. If there is a tie, you can just pick one.

You will need to divide that distance by two because half of the differences should have occurred in each of the two lineages. Begin drawing your tree by creating a branch of that length separating the two most similar sequences, in this case D and E, which only have 2 differences. This is a good time to remind students that it does not matter whether D is on top or E is on top. The interpretation is the same (**Activity 1**).



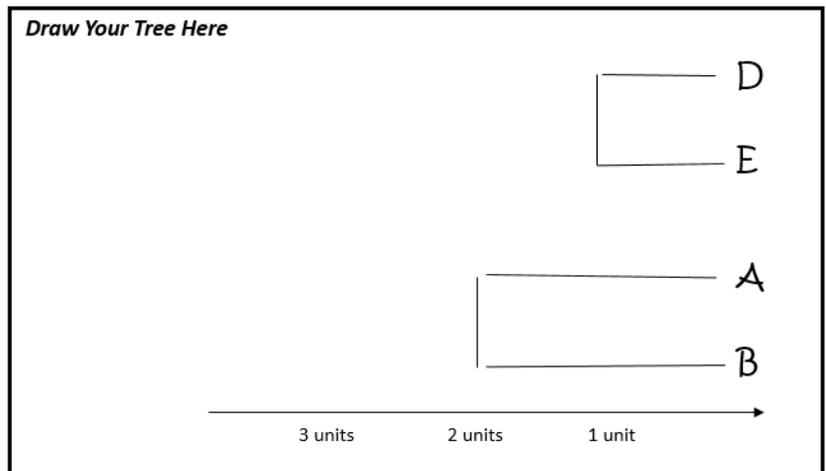
DNA Sequence Evolution Simulation and Phylogeny Building with Pen and Paper

Round 2: Next, you will collapse the matrix. First label the matrix, grouping the names of the sequences that you joined in the previous step. Then fill in the new combined row by averaging. In this case, we sum the distance between (D and A) with the distance between (E and A) then divide by two $(6+6)/2=6$ to get the distance value for the first combined cell. Now the distance corresponds to the average distance between members of the D-E clade and sequence A.

Repeat the process for the entire combined row. Then, copy the remaining distances from the previous matrix. When you are done, it should look something like the one below.

| | A | B | C | D-E |
|-----|---|---|---|-----|
| A | - | | | |
| B | | - | | |
| C | | | - | |
| D-E | 6 | | | - |

| | A | B | C | D-E |
|-----|---|---|---|-----|
| A | - | | | |
| B | 4 | - | | |
| C | 6 | 6 | - | |
| D-E | 6 | 6 | 4 | - |



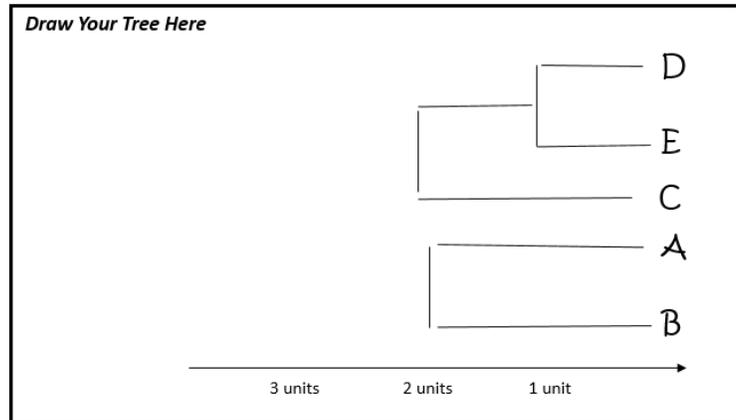
Again, find the smallest value, and draw a corresponding branch that divides the distance between the two lineages. In this case we have a tie between (A and B) and (D-E and C). Let's arbitrarily pick (A and B). The total distance is 4, we divide that between the two lineages and draw the A-B clade. Note that the A-B clade is not yet connected to the D-E clade.

Tips: The distance matrix we are working with here is "perfect", meaning that these are the distances you would expect if there were not multiple mutations at the same position of the sequence during the simulation. It is very likely that your students will not have such clean results from their simulations in **Activity 2**. This means they will not have all even numbers in the matrix, and will need to use fractional distances. It also means that they ultimately may not reconstruct a tree that matches the starting tree from the simulations. If these complexities are above and beyond what you want to tackle in this lesson, you may want to give them the initial distance matrix we use here. For more ambitious students with solid arithmetic skills, we recommend letting them forge ahead with their own data, warts and all.

DNA Sequence Evolution Simulation and Phylogeny Building with Pen and Paper

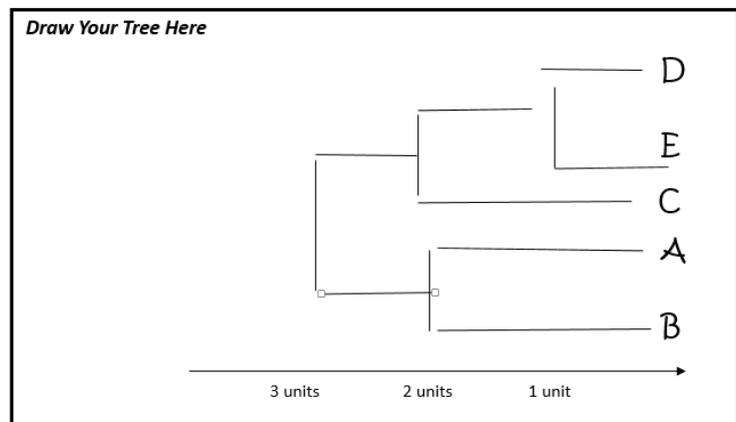
Round 3: Collapse by averaging the distances between (A and C) and (B and C) to create a combined distance between (A-B and C) and repeat to get a distance between (A-B and D-E), copy the remaining distance between (D-E and C) from the previous matrix. Find the smallest remaining value, and add the branch to your tree.

| | A-B | C | D-E |
|-----|-----|---|-----|
| A-B | - | | |
| C | 6 | - | |
| D-E | 6 | 4 | - |



Round 4: Repeat the process one more time. In this example, the final step is adding a branch with total length of 6 (3 for each lineage) to link the (C-(D-E)) clade to the (A-B) clade.

| | A-B | C-(D-E) |
|---------|-----|---------|
| A-B | - | |
| C-(D-E) | 6 | - |



Note that the final tree we obtained here is the same as the starting tree used in the simulation (**Activity 2**) even though some of the branches have rotated. This is a good opportunity to revisit the skills in phylogenetic tree interpretation from **Activity 1** to persuade students that the trees are equivalent.

If students are working with their own simulated data, ask them whether they arrived at the same answer. If not, ask them to consider why. If they need a hint, tell them to think back to both the assumptions of the model from **Activity 2** and the same reasons why it was difficult to assign the final simulated sequences back to the known tree in **Activity 3**. We used the UPGMA method to build this tree. More complicated methods that scientists typically use to analyze DNA sequence data use sophisticated statistical approaches to overcome these limitations. Still, the UPGMA method is a good way to get a quick and easy answer, and scientists often use UPGMA used to graph the relationship between other types of data, such as gene expression patterns.

Student Handout – Phylogenetic Tree

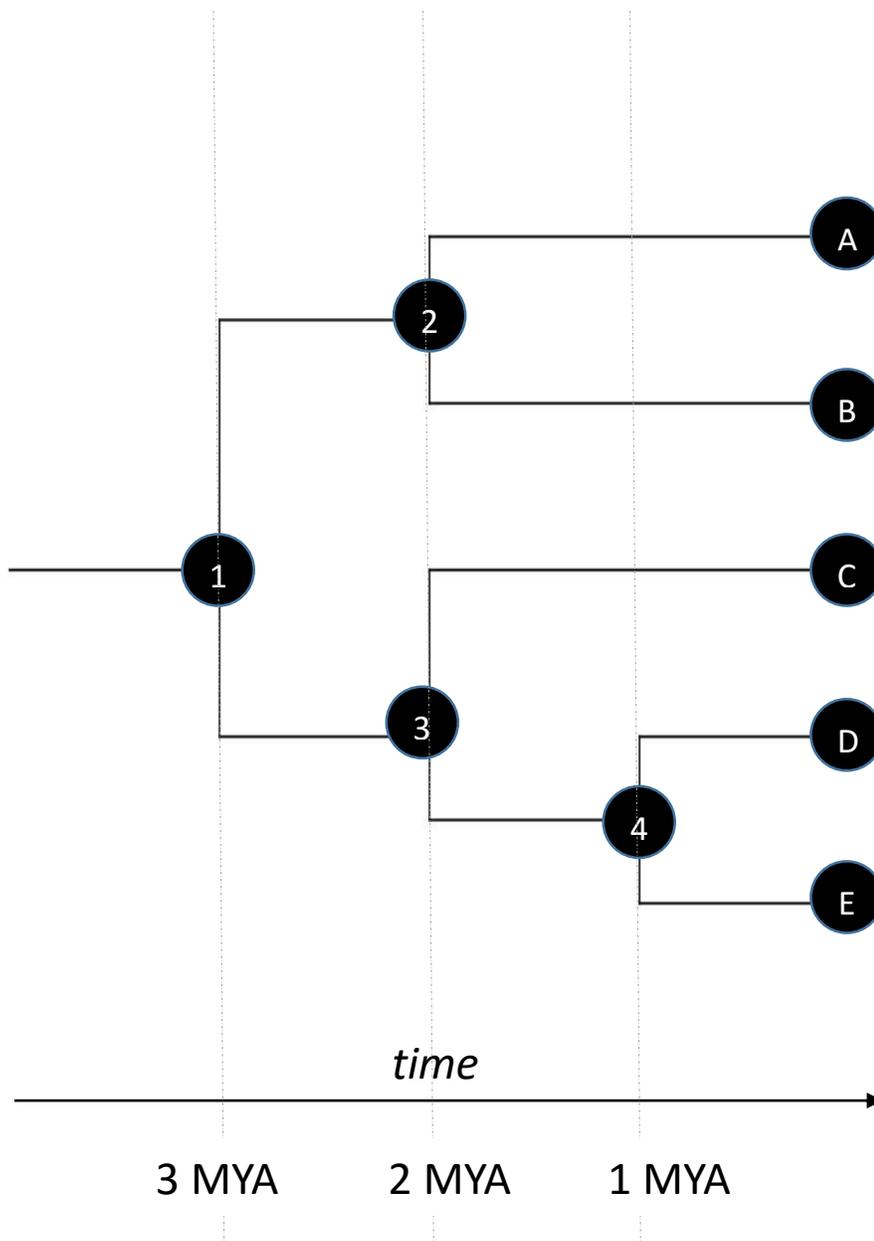


Figure 1. Phylogenetic Tree showing a model of the relationships among five organisms or genes (A-E) and four ancestral nodes (1-4). The time scale is labeled in units of millions of years ago (MYA).

Tree Thinking Worksheet

Instructions: Use the phylogenetic tree from Student Handout 1 to answer the following questions, assuming that A-E represent species.

- 1. Which species shares the most recent common ancestor with species A?**
 - a. Species B
 - b. Species C
 - c. Species D
 - d. Species E
 - e. None of the other species share a common ancestor with species A

 - 2. Species C is more closely related to Species D than to Species E.**
 - a. True
 - b. False

 - 3. Which ancestral node represents the most recent common ancestor of all the species in the phylogenetic tree?**
 - a. Node 1
 - b. Node 2
 - c. Node 3
 - d. Node 4
 - e. None of the nodes represent common ancestors

 - 4. How long has Species A been evolving independently from Species B?**
 - a. Less than 1 million years
 - b. 1 million years
 - c. 2 million years
 - d. 3 million years
 - e. Longer than 3 million years

 - 5. Homologous DNA sequences from Species C and Species D are expected to be more similar to each other than DNA sequences from Species C and Species A.**
 - a. True
 - b. False
-

Student Handout – DNA Sequence Evolution Simulation

Background:

DNA is a polymer. That is, a strand of DNA is a chain of many linked simple building blocks. There are only four different types of building blocks, or nucleotides, in a molecule of DNA. All four include the same backbone made of alternating sugars (deoxyribose) and phosphates. The only difference between the four building blocks is the part called the “base”. The four different bases are called adenine, cytosine, guanine and thymine, or A, C, G, and T, for short.

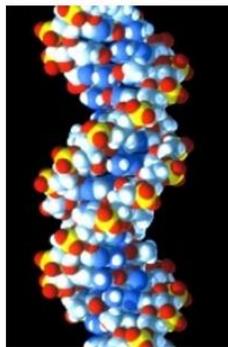
A complete molecule of DNA includes two strands of these repeating blocks twisted into a double helix. The bases on the two strands are held together by hydrogen bonds. Each base can only bind to one other type of base. A always pairs with T, and G always pairs with C. This is what allows living cells to replicate their DNA before cell division. This base pairing also means that we can describe a DNA sequence in short hand, just by specifying the order of building blocks on one of the two strands. For example, if I tell you that one strand of a DNA sequence is GGAATTCC, you automatically know the other strand is CCTTAAGG.

During replication, each strand serves as the template to make a copy of the other strand. By copying the two strands of a single molecule, a cell can produce two identical copies of the original DNA molecule...unless a mutation occurs, in which case the two new copies will be slightly different from each other. A common type of mutation replaces a single building block with one that contains a different base. For example, an A can change to a G.

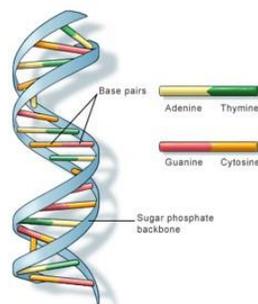
There is DNA inside of living cells

it's the molecular instructions for doing just about everything

This is a space-filling model of DNA

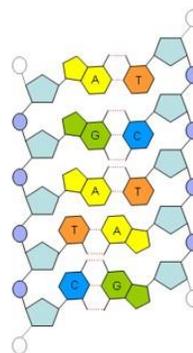


What is DNA?



U.S. National Library of Medicine

If we untwist the helix and look at the bases, we can see there are four different kinds (ACGT).



If we simplify the model a little bit, it's easier to see the double helix. There are two long strands of “backbone” with “bases” in between.

DNA Sequence Evolution Simulation and Phylogeny Building with Pen and Paper

Overview: In this exercise, you are going to simulate evolution of DNA sequences. Simulations are an effort to imitate a process, in this case, mutation of DNA sequences over evolutionary time. Simulations play an important role in science. For example, many new methods in bioinformatics (a field of inquiry dedicated to developing new computational methods to analyze biological data) are tested on simulated data sets. An advantage of working with simulated data is that many aspects of the correct answer are known, and often that is not the case with real data. Testing a new analysis method on simulated data allows scientists to make statements like “If real data arises through a natural process that is accurately imitated by this simulation, my new method of analyzing it gets the correct answer.”

Instructions: Simulate evolution starting with a 10 base ancestral DNA sequence. Roll the dice to determine which base will mutate and what type of mutation will occur. Fill in the DNA sequences for all the internal nodes and tips shown in the *DNA Simulation Worksheet*.

Each group will begin with a 10 nucleotide ancestral sequence.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| A | T | C | C | G | C | G | G | A | T |

For every million years of time that passes, you will introduce one mutation.



You will roll two dice. One has 10 sides, the other has 4 sides.



The d10 will determine which nucleotide gets the mutation. Here you rolled a “4”, so it will be the 4th nucleotide.

The d4 will determine what type of mutation will occur:

- 1=A
- 2=C
- 3=G
- 4=T

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| A | T | C | T | G | C | G | G | A | T |

Here, you rolled a 4, so the mutation will change the nucleotide to a “T”

Tips:

- If you work in groups of 3 people, one can roll the d10, one can roll the d4, and one can record the sequence changes.
- For the d10, rolling a zero means the mutation occurs in the 10th nucleotide.
- To read the d4, look at the number at the top point. It will be the same on all three visible sides of the d4.

DNA Sequence Evolution Worksheet 1

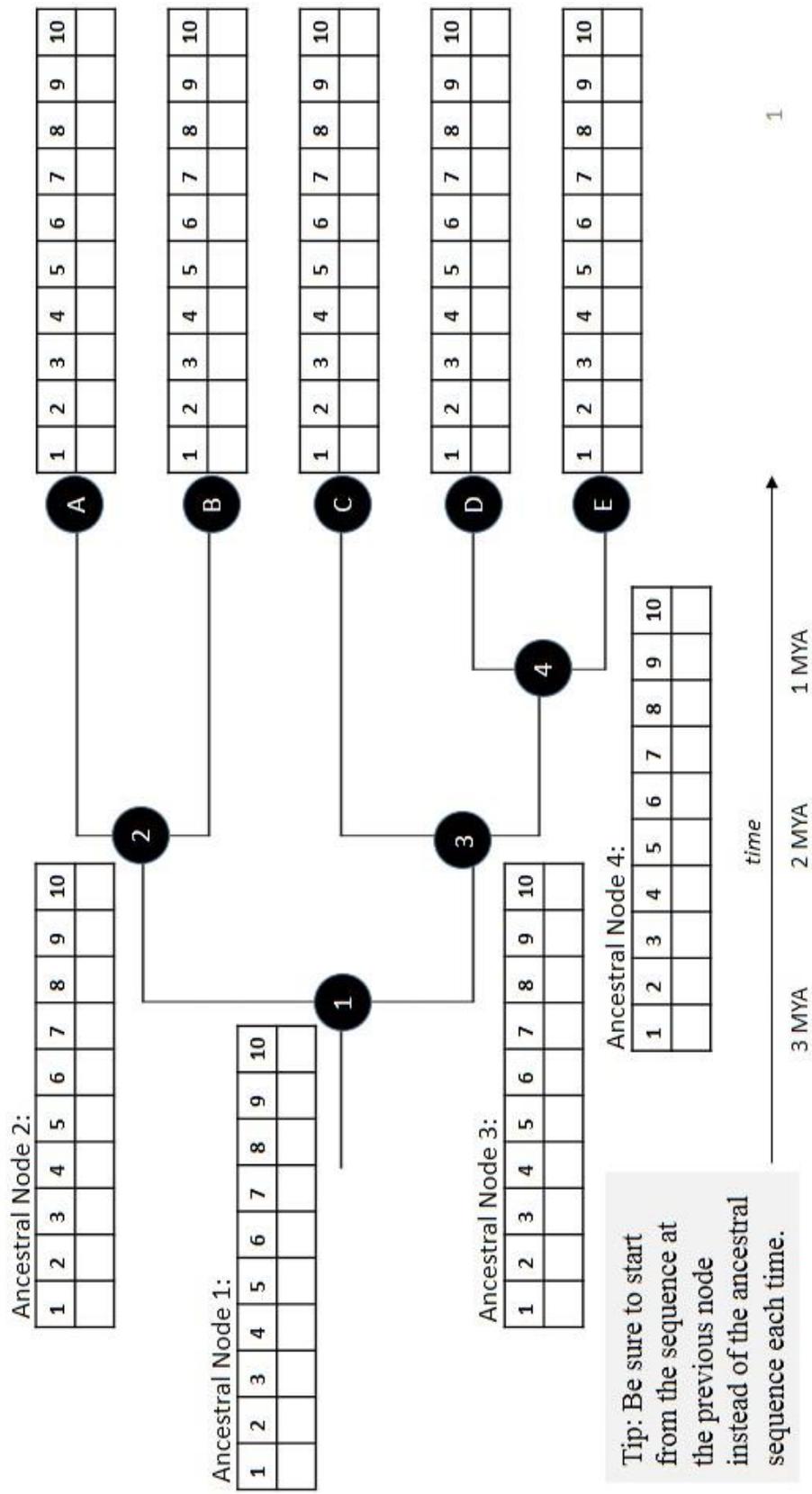
Mutations:
 1=A 3=G
 2=C 4=T

Tip: Roll the dice, record the mutation first, then fill in the rest of the sequence.

Starting Sequence:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| A | T | C | C | G | C | G | G | A | T |

Tip: Pay attention to the branch length. A branch two million years long should get two mutations.



Tip: Be sure to start from the sequence at the previous node instead of the ancestral sequence each time.

DNA Sequence Evolution Worksheet 2

Cut along the dotted lines.

time
3 MYA 2 MYA 1 MYA

Code the leaf node labels to disguise the order

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|----|

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|----|

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|----|

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|----|

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|----|

UPGMA Worksheet

Round 1: Count the number of differences between each pair of sequences

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | - | | | | |
| B | | - | | | |
| C | | | - | | |
| D | | | | - | |
| E | | | | | - |

Round 3: Collapse by averaging distances to the most similar pair

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | - | | | | |
| B | | - | | | |
| C | | | - | | |
| D | | | | - | |
| E | | | | | - |

Round 2: Collapse the matrix by averaging distances to the most similar pair

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | - | | | | |
| B | | - | | | |
| C | | | - | | |
| D | | | | - | |
| E | | | | | - |

Round 4: Collapse by averaging distances to the most similar pair

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | - | | | | |
| B | | - | | | |
| C | | | - | | |
| D | | | | - | |
| E | | | | | - |

Draw Your Tree Here